PA153 Počítačové zpracování přirozeného jazyka

11 – Znalosti, parafráze, odvozování

Karel Pala, Marek Medveď

Centrum ZPJ, FI MU, Brno

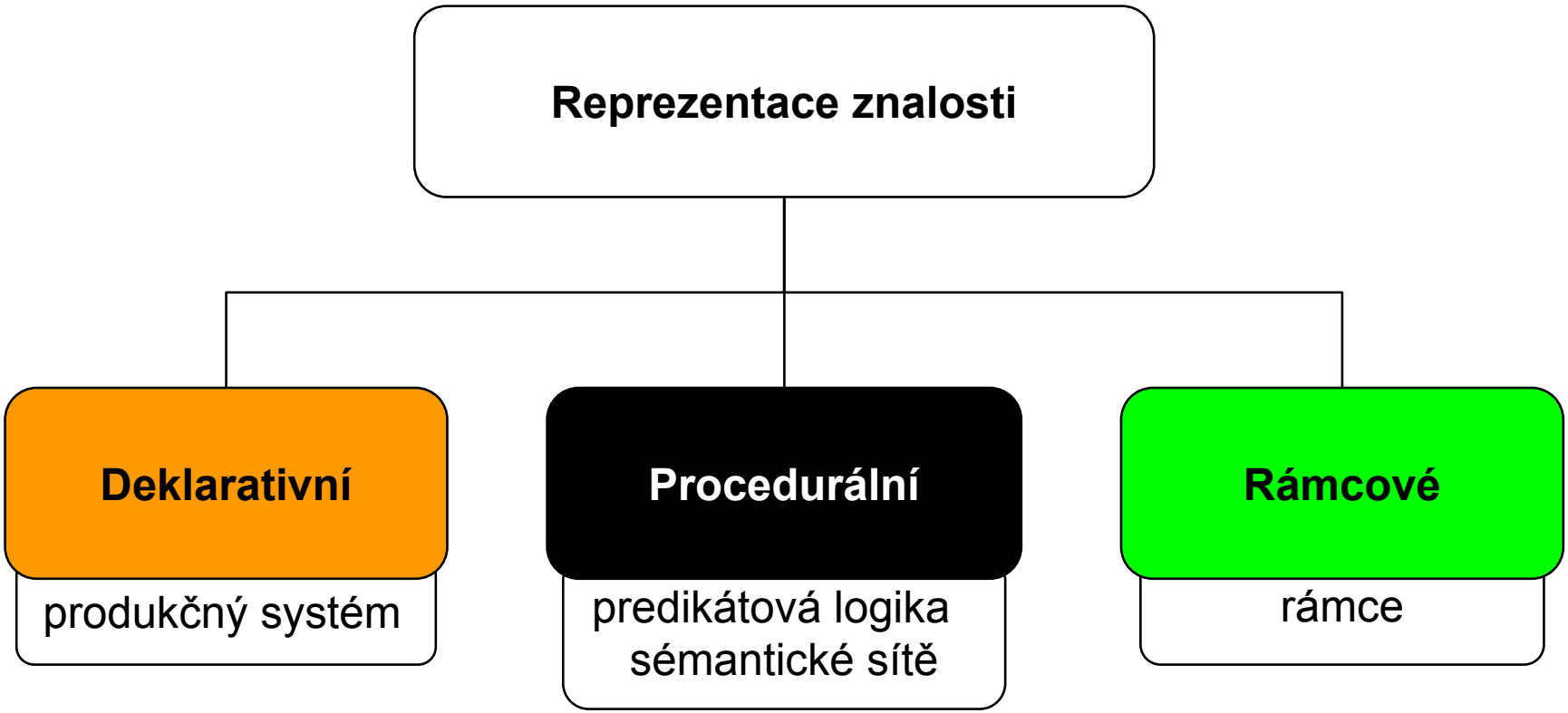
5. prosince 2018

Znalosti

- znalosti o jazyce (lexikon, gramatické kategorie, syntax)
- znalosti o světě

Znalostní báze (knowledge base, KB): obsahuje fakta, která jsou premisami v deduktivním odvozování
lidmi čitelné KB: how-to, FAQ, recepty, návody, diagramy
strojově čitelné KB: ontologie (SUMO-MILO), sémantické sítě (WordNet), dbPedia

Reprezentace znalosti



Znalosti

Deklarativní:

- znalosti zaznamenané v určitém jazyce
- uložené v určitém zdroji (databáza)
- jednoduché odvozován
- explicitná
- formálně verifikovatelná
- obecně platná

Procedurální:

- vyjádření pomocí procedury
- hodnota se zjistí provedením procedury
- implicitní

Example (pohyb robota po místnosti)

Deklarativní: pohyb robota + mapa
Procedurální: príď na pozíciu (X,Y)

Znalosti

Rámce

- kombinace deklarativního a procedurálního přístupu
- rámce samotné sú deklaratívne
- sloty v rámcoch sú procedurálne

Odvozování

Reprezentace znalostí (knowledge representation): znalostní báze + odvozovací pravidla

Dva druhy:

- deduktivní odvozování
- nededuktivní odvozování

Deduktivní odvozování: monotónní a nemonotónní odvozování [Allen, 1995]

KB: Ptáci létají. Vrabec je pták.

Vrabec létá.

Deduktivní odvozování: monotónní a nemonotónní odvozování [Allen, 1995]

KB: Ptáci létají. Vrabec je pták. Pštros je pták.

Vrabec létá. Pštros létá.

Deduktivní odvozování: monotónní a nemonotónní odvozování [Allen, 1995]

KB: Ptáci létají. Vrabec je pták. Pštros je pták. Pštros nelétá.

Vrabec létá. ~~Pštros létá.~~

Znalosti o světě

- encyklopedické (Jaké je hlavní město ČR?)
- common-sense (Jak je vhodné obléci se 5. prosince 2018?)

počítačově zpracovatelné zdroje encyklopedických znalostí:

- encyklopedie
- znalostní hry
- dbPedia: strojově zpracovaná Wikipedie

Common sense a odvozování

common sense: sdílená znalost, ne vždy v souladu s (vědeckými) fakty (V noci nesvítí slunce.)

Cheap apartments are rare.
Rare things are expensive.

Cheap apartments are expensive.

Deduktivní odvozování není možné použít vždy (ve skutečnosti skoro nikdy).

Common sense: nejznámější projekty

- Never-ending Language Learning (NELL):
  - prochází web a odvozuje (hledá spojení mezi věcmi, které zná a věcmi, které najde prostřednictvím vyhledávání)
  - pr. Pikes Peak
  - občas nutný lidský zásah ("I deleted my (Internet) cookies", "I deleted my files" $\Rightarrow$ soubor je stejná kategorie jako pečivo)
- CyC: vývoj od r. 1985(!)
  - reprezentace pomocí vlastního jazyka CyCL
  - pokus o zavedení obsáhlé ontologie a znalostní báze
  - cíl: expresivní jazyk, ontologie v rozumné úrovni detailu, znalostní báze, rychlý inferenční systém
  - ontologia: 1,5 M tokenov
  - KB: 24,5 M pravidiel
  - inferenčný systém: dedukcia, indukcia, machine learning
- ConceptNet: syntaktická analýza OpenMind, propojení s Wiktionary

Parafráze

Parafráze: promluva x je parafrází promluvy y, pokud x a y mají stejný nebo podobný význam.

Tento most postavila Nejlepší firma s.r.o.
Nejlepší firma s.r.o. postavila tento most.
Stavitelem tohoto mostu je Nejlepší firma s.r.o.

Přesnější definice

Textové vyplývání $\neq$ logické vyplývání

Z text t textově vyplývá hypotéza h ($t \Rightarrow h$), pokud lidé, kteří přečtou t, odvodí, že h je nejspíš pravda. [Dagan et al., 2007]
parafráze $= h \Rightarrow t \wedge t \Rightarrow h$

Rozpoznávání textových vyplývání/parafrází

hledání podobností:

- na řetězcích (např. Levenshteinova vzdálenost)
- na slovech
- na slovech s použitím znalostní báze (např. slovník synonym)
- na syntaktických stromech
- kombinace předchozích

Rozpoznávání textových vyplývání/parafrází

využití:

- odpovídání na otázky
- chatbots
- detekce plagiátů
- výuka
- automatická sumarizace textu
- doplnění implicitní znalosti
  - logická analýza textu
  - znalostní modely v umělé inteligenci
- ...

Korpusy parafrází

- Microsoft Research Paraphrase Corpus[1]
- The Boeing-Princeton-ISI (BPI) Textual Entailment Test Suite[2]
- Multiple Translation Chinese Corpus[3]
- The SEMILAR Corpus: The SEMantic SimILARity Corpus[4]
- Paraphrase Discovery[5]

[1] http://research.microsoft.com/en-us/downloads/607d14d9-20cd-47e3-85bc-a2f65cd28042/

[2] http://www.cs.utexas.edu/users/pclark/bpi-test-suite/

[3] https://catalog.ldc.upenn.edu/LDC2002T01

[4] http://deeptutor2.memphis.edu/Semilar-Web/public/semilar-api.html

[5] http://nlp.cs.nyu.edu/paraphrase/

Paraphrase Discovery

vztahy mezi pojmenovanými entitami v korpusových datech:

```
[lemma="Hannibal"] []* [lemma="Hopkins"] within <s/>
```

ztvárnit	jako
hrát	odmítnout
s	na roli
si	hrající
/	se objevil
v podání	představoval
alias	působí v roli
se svým přítelem	
(	
po boku	

Generování parafrází

Základní způsoby parafrázování:

- aktivní–pasivní větná konstrukce: Tento most byl postaven Nejlepší firmou s.r.o.
- synonyma: Tuto lávku postavila Nejlepší firma s.r.o.
- hyperonyma: Tuto stavbu postavila Nejlepší firma s.r.o.
- substantivizace, deverbalizace: Stavitelem tohoto mostu je Nejlepší firma s.r.o.
- kombinace: Tento most byl vytvořen Nejlepší firmou s.r.o.

Podrobněji v [Bhagat and Hovy, 2013].

Přirozená logika [Lakoff, 1970]

nástrojem této logiky je přirozený jazyk

- monotonicita (monotonicity): víc než tisíc je hodně
  Mám víc než tisíc knih. Mám hodně knih.
  Nemám víc než tisíc knih. Nemám hodně knih.
- obsažení/omezení (containment): červené auto je auto
  Po ulici jelo červené auto. Po ulici jelo auto.
  Po ulici nejelo červené auto. Po ulici nejelo auto.
- exkluze (exclusion): pes není kočka
  Na dvorku seděl pes. Na dvorku seděla kočka.
  Na dvorku neseděl pes. Na dvorku neseděla kočka.

odvození vs. presupozice (podprahové informace):
Mark David Chapman zastřelil Johna Lennona. $\Rightarrow$ John Lennon nežije.

Brazílie vyhrála mistrovství světa.$\Rightarrow$ Brazílie hrála na mistrovství světa.

Přirozená logika [Lakoff, 1970]

nástrojem této logiky je přirozený jazyk

- monotonicita (monotonicity): víc než tisíc je hodně
  Mám víc než tisíc knih. Mám hodně knih.
  Nemám víc než tisíc knih. Nemám hodně knih.
- obsažení/omezení (containment): červené auto je auto
  Po ulici jelo červené auto. Po ulici jelo auto.
  Po ulici nejelo červené auto. Po ulici nejelo auto.
- exkluze (exclusion): pes není kočka
  Na dvorku seděl pes. Na dvorku seděla kočka.
  Na dvorku neseděl pes. Na dvorku neseděla kočka.

odvození vs. presupozice (podprahové informace):
Mark David Chapman zastřelil Johna Lennona. $\Rightarrow$ John Lennon nežije.
Mark David Chapman nezastřelil Johna Lennona. $\not\Rightarrow$ John Lennon nežije.
Brazílie vyhrála mistrovství světa.$\Rightarrow$ Brazílie hrála na mistrovství světa.
Brazílie nevyhrála mistrovství světa.$\Rightarrow$ Brazílie hrála na mistrovství světa.

BDI: Znalost nebo domněnka?

KB: Ptáci létají. Vrabec je pták. Pštros je pták. Pštros nelétá. Mrtvý vrabec nelétá.
Znalostní báze se mění. Některé znalosti mají poměrně krátké trvání (Nejsem unavená. Je půl čtvrté.)
V umělé inteligenci se používá termín domněnka (belief) [Mařík et al., 2001].

Umělá inteligence: modely uvažování inteligentních agentů

Intencionální systém: agent umí ,,uvažovat“ o svých znalostech. Je schopen přemýšlet o svých přáních a jak jich lze dosáhnout [Mařík et al., 2001].

Mentální postoje:

- informační postoje – znalosti, fakta získaná senzory
- proaktivní postoje – cíle, plány, závazky

Psychologické modely lidského uvažování [Bratman, 1987]: kognitivní stavy, afektivní stavy, konnativní stavy.

Domněnka–přání–záměr: softwarový model pro aktivní inteligentní agenty

Umělá inteligence: belief–desire–intention

Záměr, Intention

Aby bylo možné vytvořit aktivního agenta, je třeba, aby ,,věděl, co chce“ (intention). Pokud ví, co chce (tj. má záměr), vytvoří si agent nějaký plán (lokální cíl).

Příklad: najdi cestu z domu X na FI

Int a ϕ agent si vybírá vždy cesty tak, aby na nich někdy platila ϕ

Přání, Desire

Přání vyjadřuje agentovu motivaci. Motivovaný agent má cíle (cílové stavy). Cíle by neměly být v rozporu.

Příklad: najdi nejkratší cestu z domu X na FI

Des a ϕ pravdivost formule ϕ je cílem agenta *a*

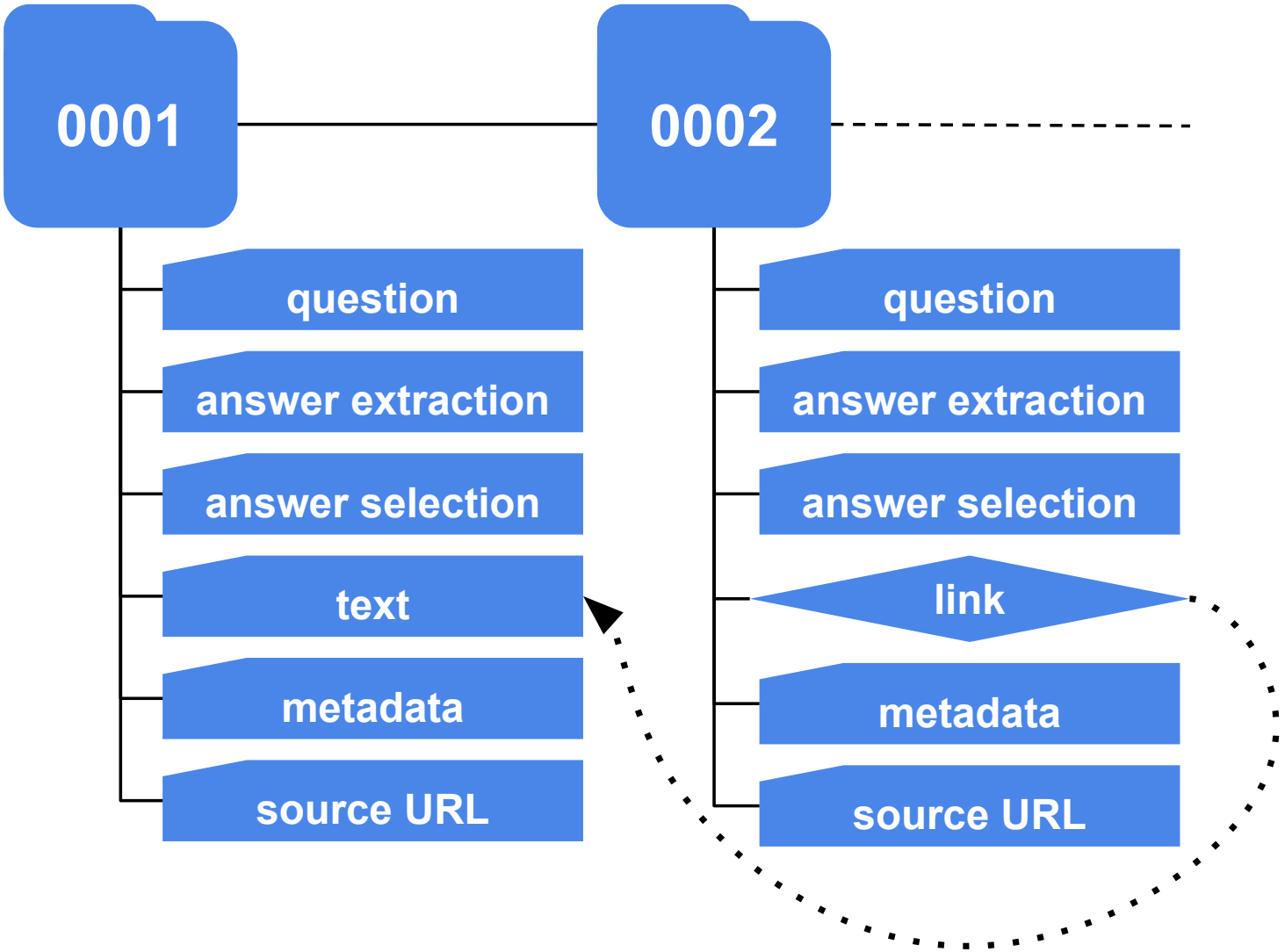
Belief, Domněnka

Domněnka představuje agentovu bázi znalostí. Informace mohou být pravdivé, agent v ně v daný okamžik věří a chápe je jako nedokonalé přiblížení obrazu okolního světa [Mařík et al., 2001].

Příklad: najdi nejkratší cestu z domu na FI. Mostecká je neprůjezdná.

Bel a ϕ agent *a* věří v pravdivost formule ϕ

Databáze SQAD

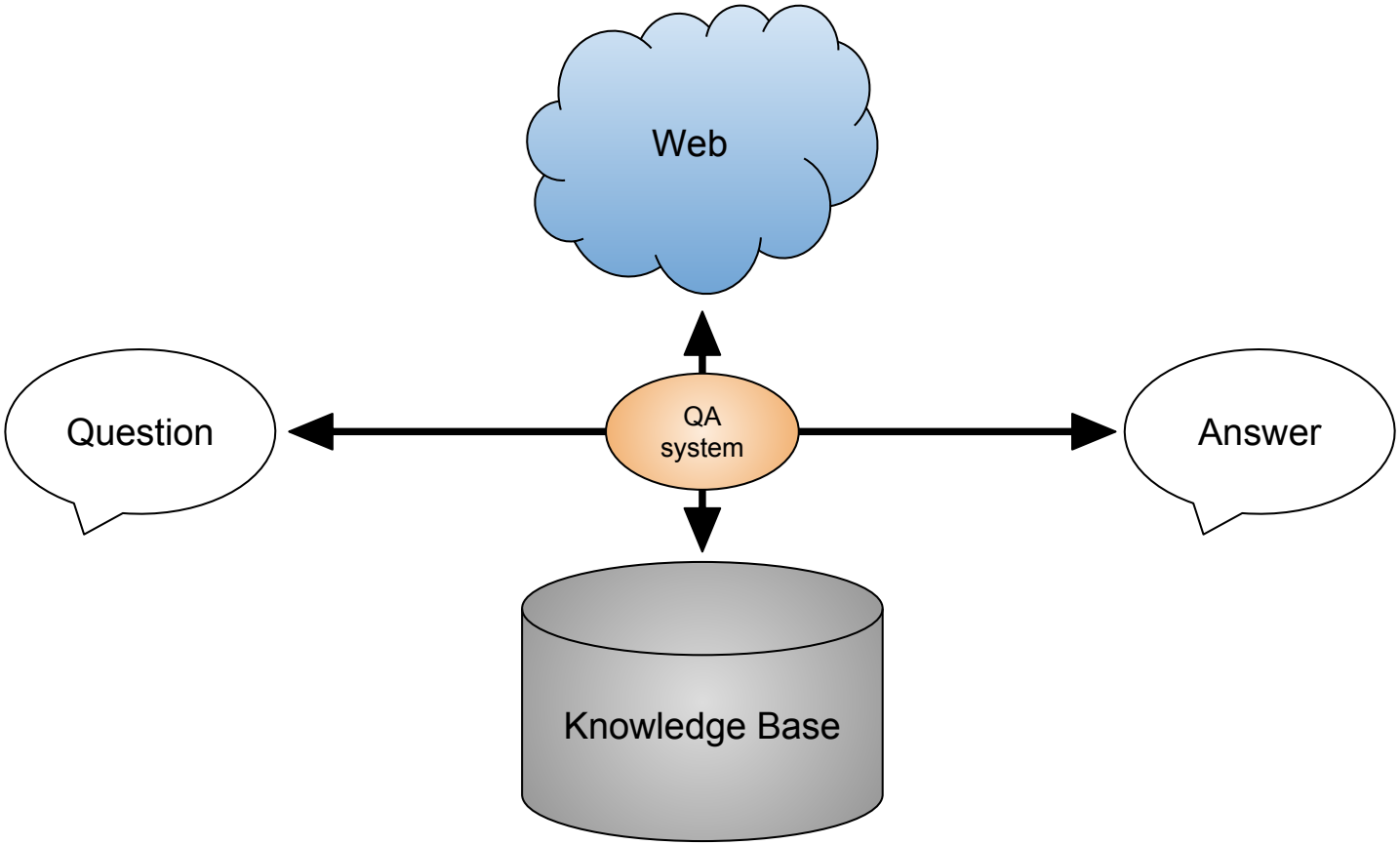


Otázka

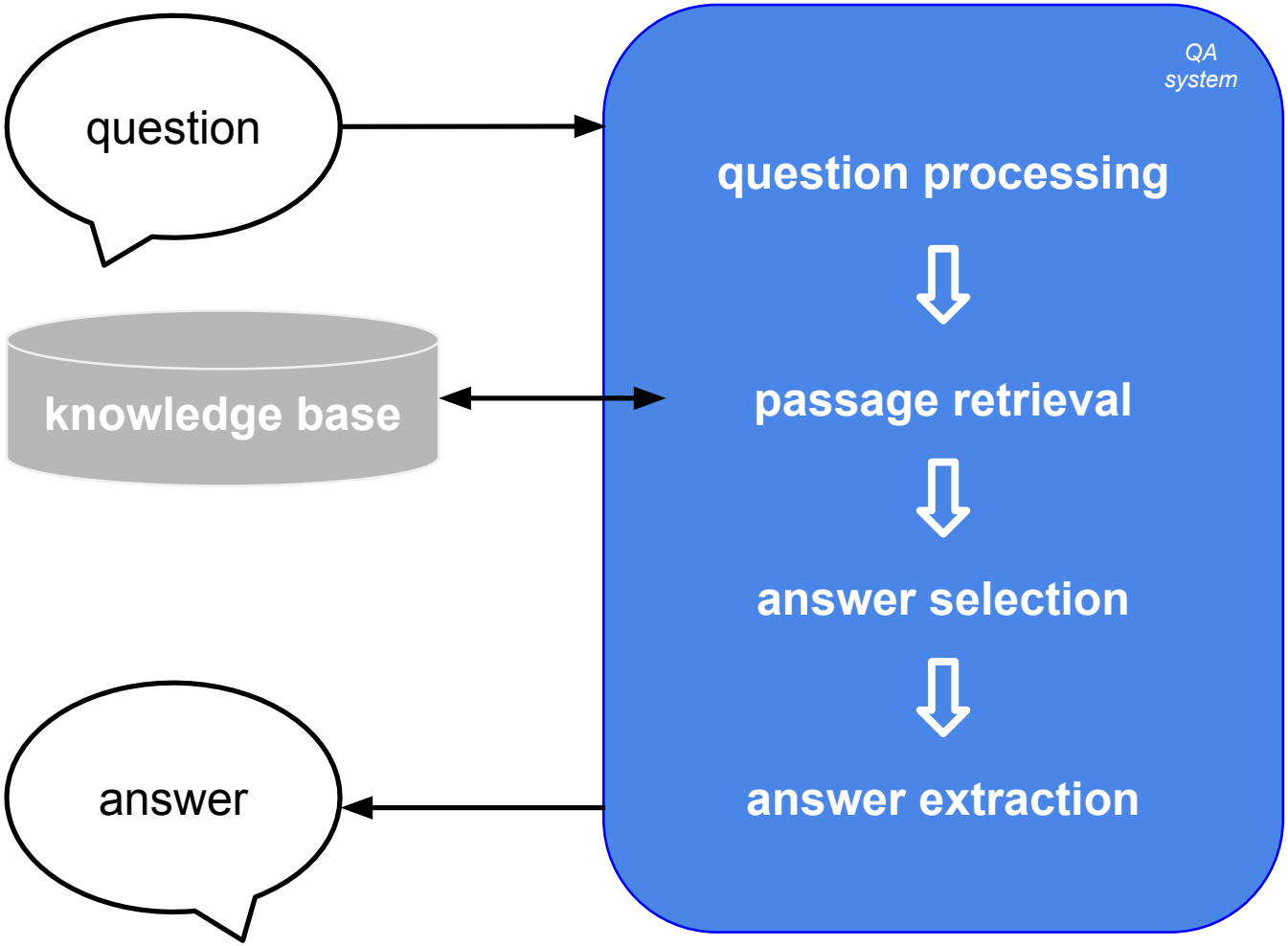
word/token	*lemma*	*tag*
<s>		
Jak	jak	k6eAd1
se	sebe	k3xPyFc4
jmenuje	jmenovat	k5eAaImIp3nS
světově	světově	k6eAd1
nejrozšířenější	rozšířený	k2eAgFnSc1d3
hra	hra	k1gFnSc1
na	na	k7c4
hrdiny	hrdina	k1gMnPc4
<g/>		
?	?	kIx.
</s>		

Odpověď: `Dungeons & Dragons`
Text: `Nejrozšířenější světově hranou RPG hrou na hrdiny pak je Dungeons & Dragons.`
Metadata: (Entity, Entity)

Question answering system



AQA


question
knowledge base
answer
QA system
question processing
passage retrieval
answer selection
answer extraction

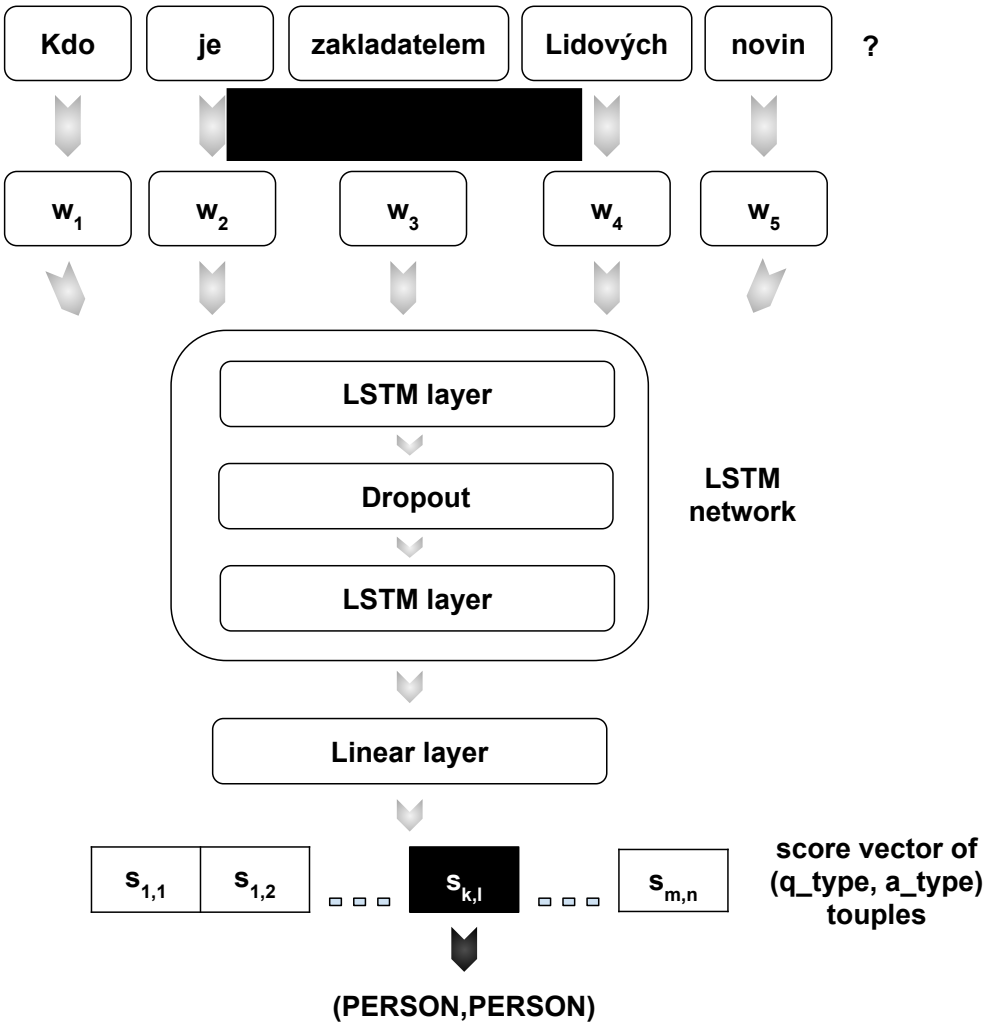
Reprezentace znalostí

- *Question reformulation*: "*Jak se jmenuje ... osoba ...*" reformuluje na "*Kdo je ...*"
- *syntactic tree*:

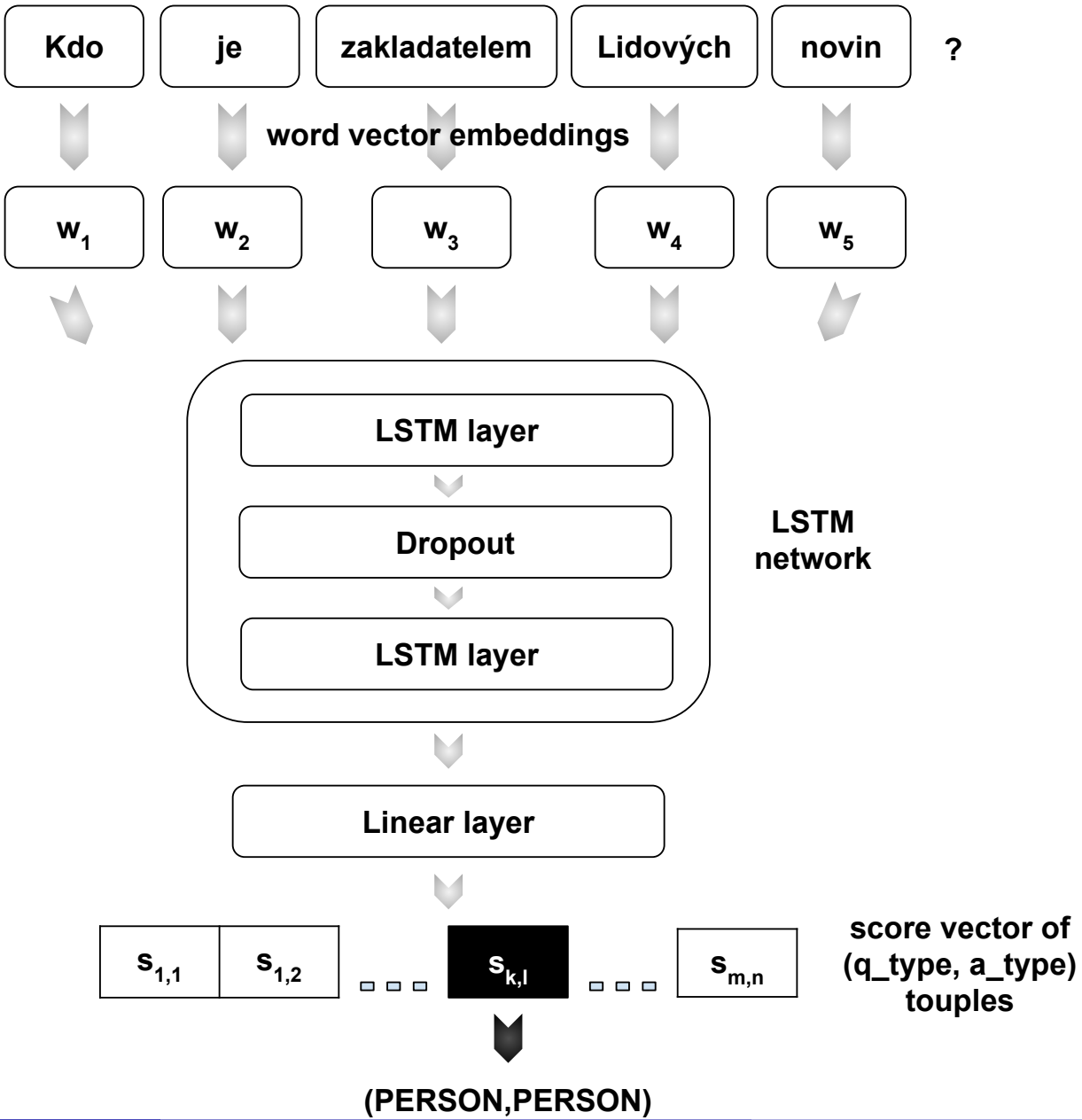
ID	*word*	*Dep ID*
0	Jak	2
1	se	2
2	jmenuje	-1
3	otec spisovatele Jiřího Mouchy	2

- *Question type extraction*: "*Kdo byl ...*" typu WHO
- *Main subject and main verb extraction*: Jak se jmenuje otec ... -¿ jmenuje (hlavní sloveso)

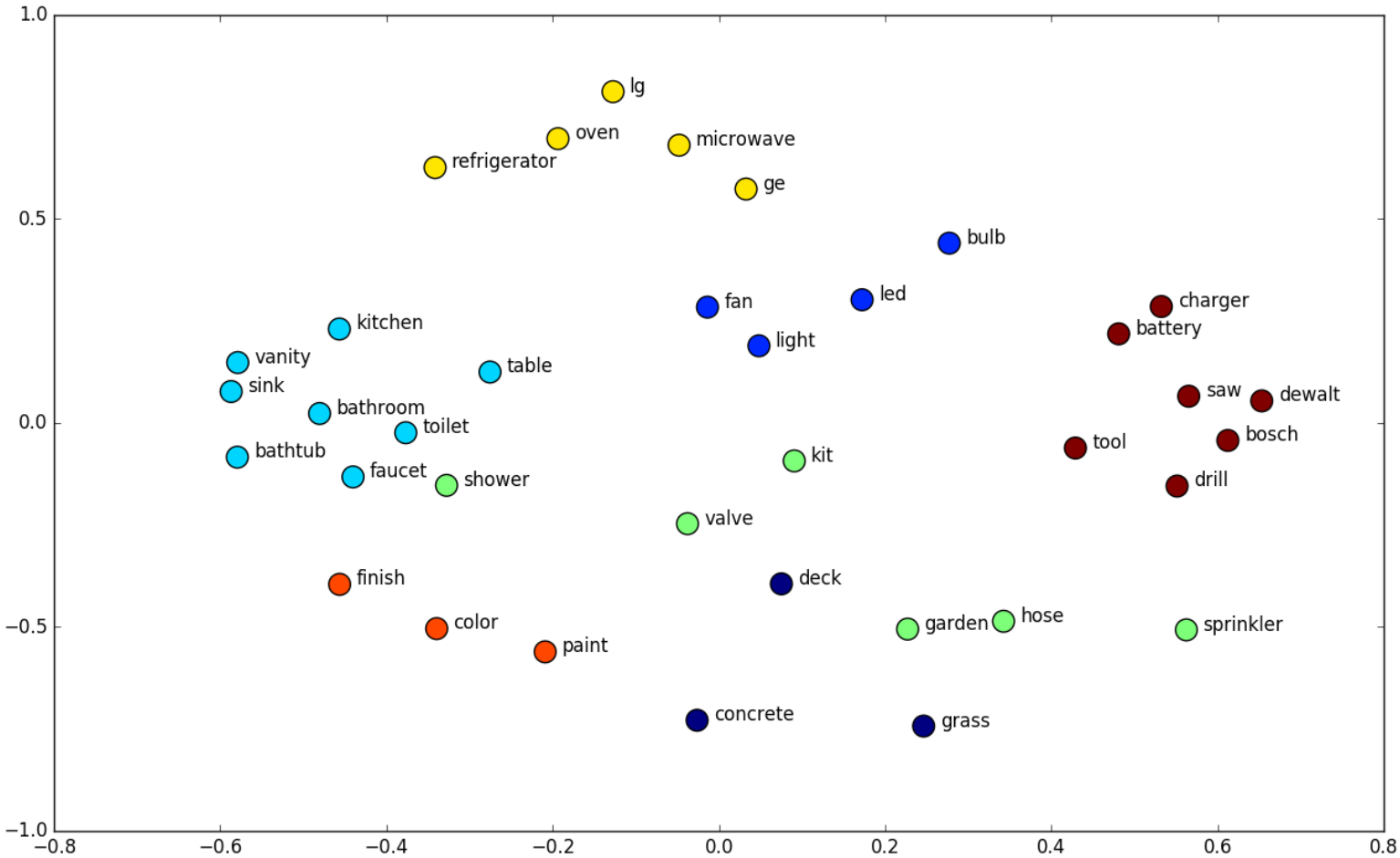
Analýza dotazů (ML based)



Analýza dotazů (ML based)



Word2vec


lg
oven
microwave
refrigerator
ge
bulb
fan
led
charger
kitchen
battery
light
vanity
table
sink
saw
dewalt
bathroom
toilet
bosch
tool
bathtub
kit
faucet
shower
drill
valve
finish
deck
color
garden
hose
sprinkler
paint
concrete
grass
1.0
0.5
0.0
−0.5
−1.0
−0.8
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8

Hyperonyma

question:	'Jak se jmenovala první manželka Miloše Formana?' (What was the name of the first wife of Miloš Forman?)
keyword:	'manželka' (wife)
hypernyms:	['manželka', 'jednotlivec', 'osoba', 'bytost', 'organismus'] (wife, individual, person, being, organism)
rule:	`(PERSON; PERSON) -> "osoba"in keyword.hypernym`

Odkazy I


Allen, J. (1995).
Natural Language Understanding (2nd ed.).
Benjamin-Cummings Publishing Co., Inc., Redwood City, CA, USA.


Bhagat, R. and Hovy, E. (2013).
What is a paraphrase?
Computational Linguistics, 39(3):463–472.


Bratman, M. (1987).
Intention, plans, and practical reason.
Harvard University Press.


Dagan, I., Roth, D., and Zanzotto, F. (2007).
Tutorial notes.
In *45th Annual Meeting of the Association of Computational Linguistics. The Association of Computational Linguistics*.

Odkazy II


Lakoff, G. (1970).
Linguistics and natural logic.
Synthese, 22(1-2):151–271.


Mařík, V., Štěpánková, O., and Lažanský, J. (2001).
Umělá inteligence.
Number svazek 3 in Umělá inteligence. Academia.